Latent rhythm transformation of drum recordings

Jason Hockman[1] *and* **Jake Drysdale**[2]

[1]*SODA*, Manchester Metropolitan University, Manchester, United Kingdom
[2]*Beatoven.ai*, Bengaluru, Karnataka, India

Correspondence Address: *j.hockman@mmu.ac.uk*

Audio Engineering Society

Overview

Redrumming: Replace or layer recorded drums with alternative recordings while preserving the original classes and timing of a target performance.

Motivation: Lightweight redrumming tool for producers to achieve rhythmic transformations not feasible through manual separation in drum recordings with layered events.

Method

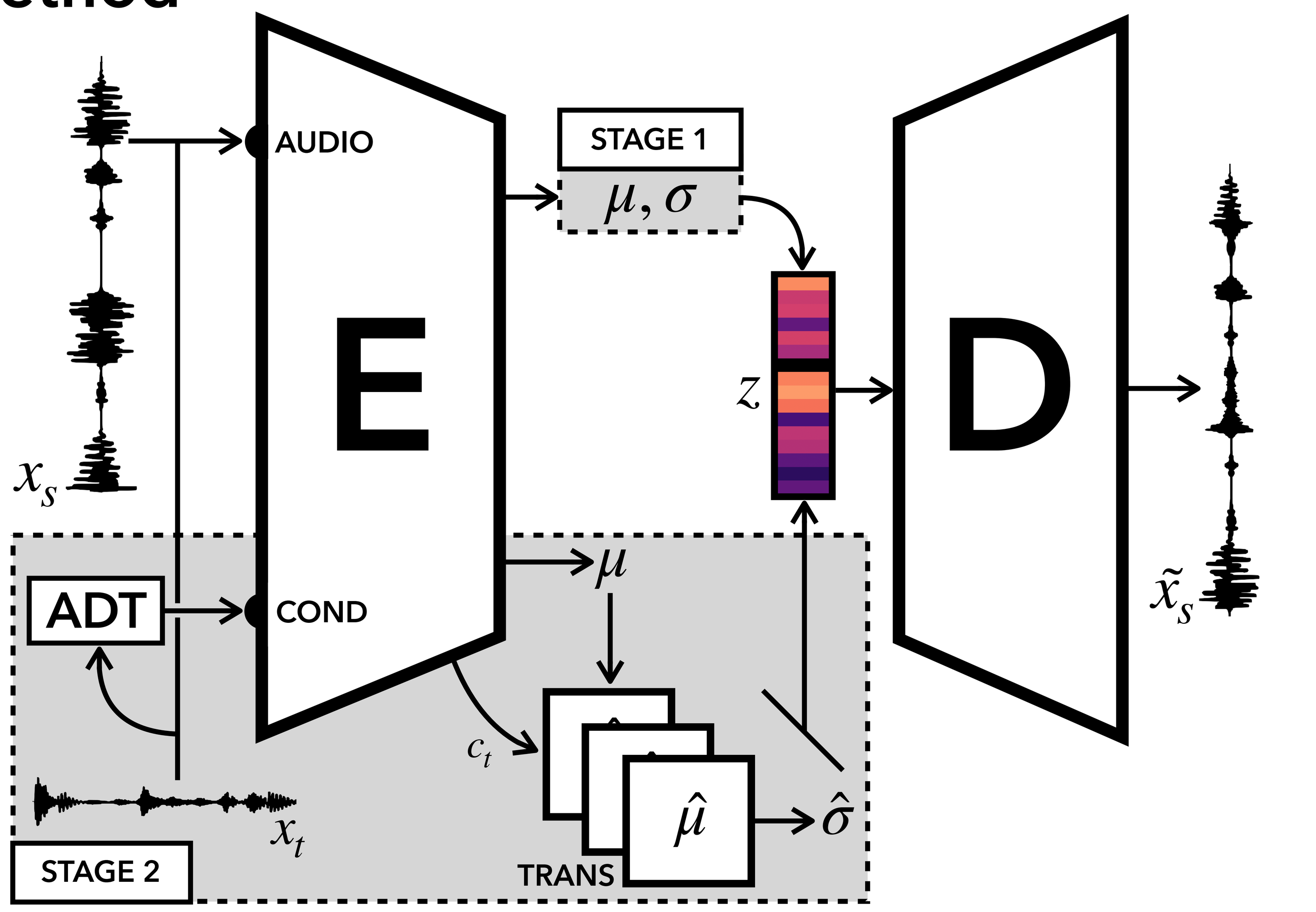

Figure 1: Model overview.

Extend RAVE to transform rhythmic characteristics (i.e., drum classes and timing) of *source* recording x_s to match those of a *target* recording x_t (*Figures 1 and 2*).

2-Stage Process:

- *Stage 1 [VAE]*: RAVE training procedure with V2 Encoder (*E*) / Decoder (*D*) and continuous latent *z* parameterised by mean μ and std σ.
- *Stage 2 [TRANSFORMER]*: Discard σ from *E*; learn rhythm-attended $\hat{\mu}$ from μ via lightweight transformer stack with conditioning c_t from target ADT probabilities and *E* timbre embedding; $\hat{\sigma}$ learned from $\hat{\mu}$ with gated residual `std_head`; $\hat{\mu}$ and $\hat{\sigma}$ used to create *z* input to *D*.

Transformer Stack:

- Style transfer formulation: *Q* = conditioning and *K*/*V* = μ.
- 3-layer stack; 4 heads each.
- Conditional embeddings projected via 4-layer Conv1D stacks (LeakyReLU; kernel 5 and 3), merged with learnable gate.
- Relative positional bias and rhythm bias from linear projection of ADT probabilities.
- μ updated via linear interpolation at each layer with learnable α.

Gated Residual `std_head`:

- 1x1 conv + GELU for channel mixing at each timestep; depthwise conv (kernel=5) for per-channel temporal context; pointwise conv for cross-channel fusing.
- Sigmoid gate mixes residual and update.

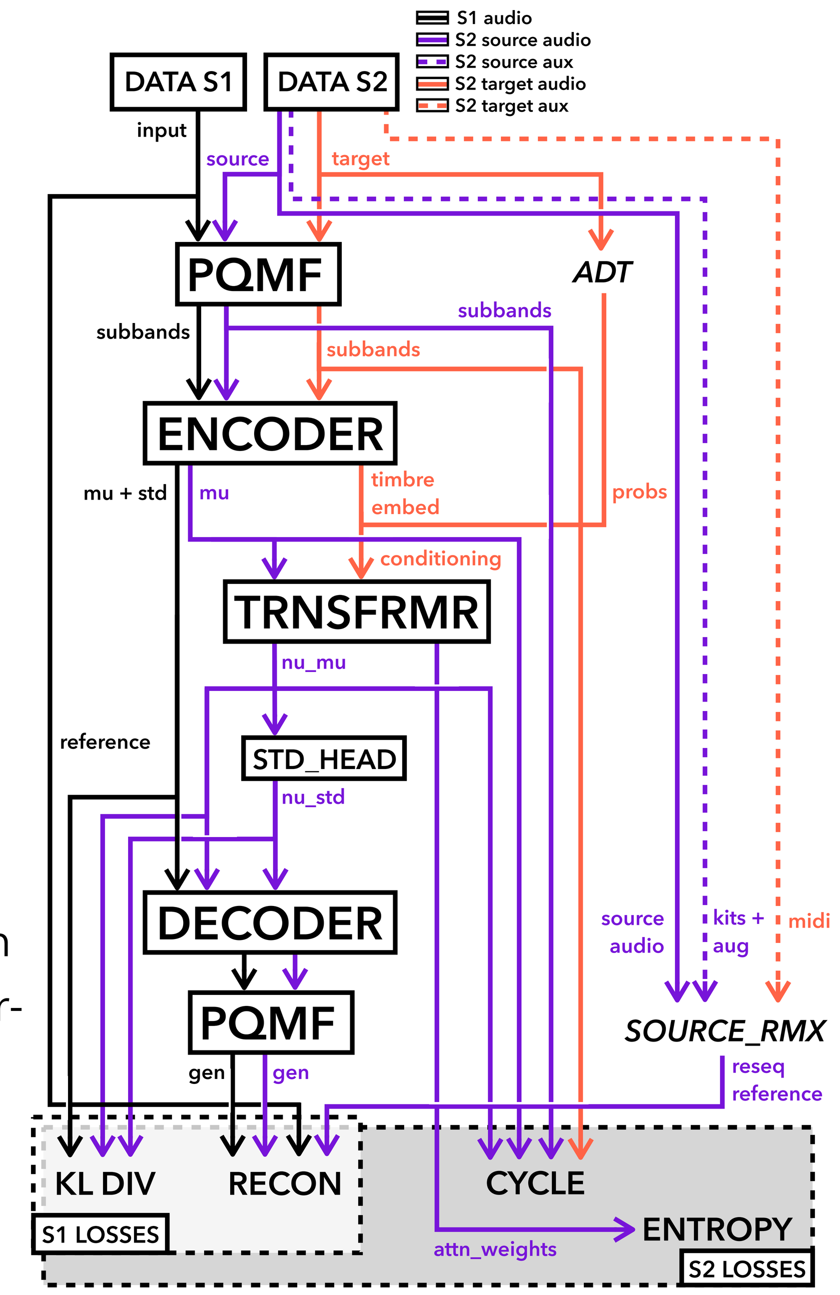

Figure 2: System modules with Stage 1 real data + Stage 2 synthetic data flow.

Training

Data:

- *Stage 1*: 10K 3-sec real drum recordings (breakbeats); randomised mute, gain, compression, dequantise, pitch-shifting and cropping during training.
- *Stage 2*: 10K 3-sec synthetic recordings made from 100 kits built from real and electronic oneshots and 100 rhythms; ADT probs, event timing and kits stored.
  - **Data creation augmentation**: Oneshot pitch and gain; tempo-scaling, swing, microtiming, and event-timing dropouts.
  - **Training augmentation**: Same as Stage 1 without cropping.
  - **Transform pairing**: Randomised and reseeded per epoch.

Losses:

- *Stage 1*: Rep Learning: 1M / Adv Training: 2M steps. **KL** + **Recon** (RAVE procedure).
- *Stage 2*: Rep+Transform Learning: 200K / Adv Training: 160K steps.
  - *z* re-regulated with **KL** to Stage 1 β_{max}=0.05 (6000 steps).
  - `source_rmx`: Standard recon unsuitable; pseudo-targets $x_{s \leftarrow t}$ built from augmented oneshots from x_s aligned to rhythm of x_t with timing params (swing, microtiming, event dropouts); **Recon** (w_r=0.1) computed against $x_{s \leftarrow t}$.
  - **Cycle consistency** (w_c=0.3) promotes invertibility.
  - **Attention entropy** (w_e=0.1) encourages head diversity.

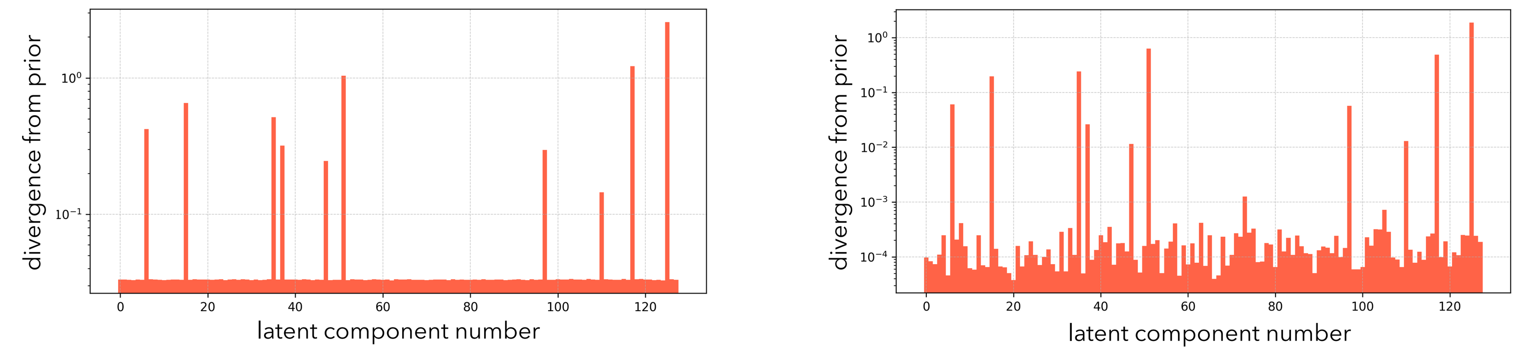

Figure 3: Mean KL per latent between posterior distribution estimated over Stage 1 (left) and Stage 2 (right) datasets and prior.

- Stage 2 leverages frozen VAE from Stage 1 (w/o σ) to produce $\hat{\mu}$ and $\hat{\sigma}$ in similar latent representational space with similar KL values (*Figure 3*).
- Stage 2 adversarial stage adapts Decoder to new parameters.
- After training, transform performed on real drums with ADT probs.

System Output and Future Work

- Output coherently conveys target rhythm, applied as intended to source timbres. (*Figure 4*).
- Where timbre not convincingly achieved in generations, traces of source are heard with reduced high-frequency detail.
- Stage 2 utilises only synthetic data; real and synthetic data exist on same manifold from VAE training; experiments indicate transform preserves timbre mapping fairly well. Further testing with larger ratio of real oneshots and additional augmentation (e.g., reverb) may improve mapping and afford more informative evaluation.
- Experimentation with loss balancing may yield further improvements.

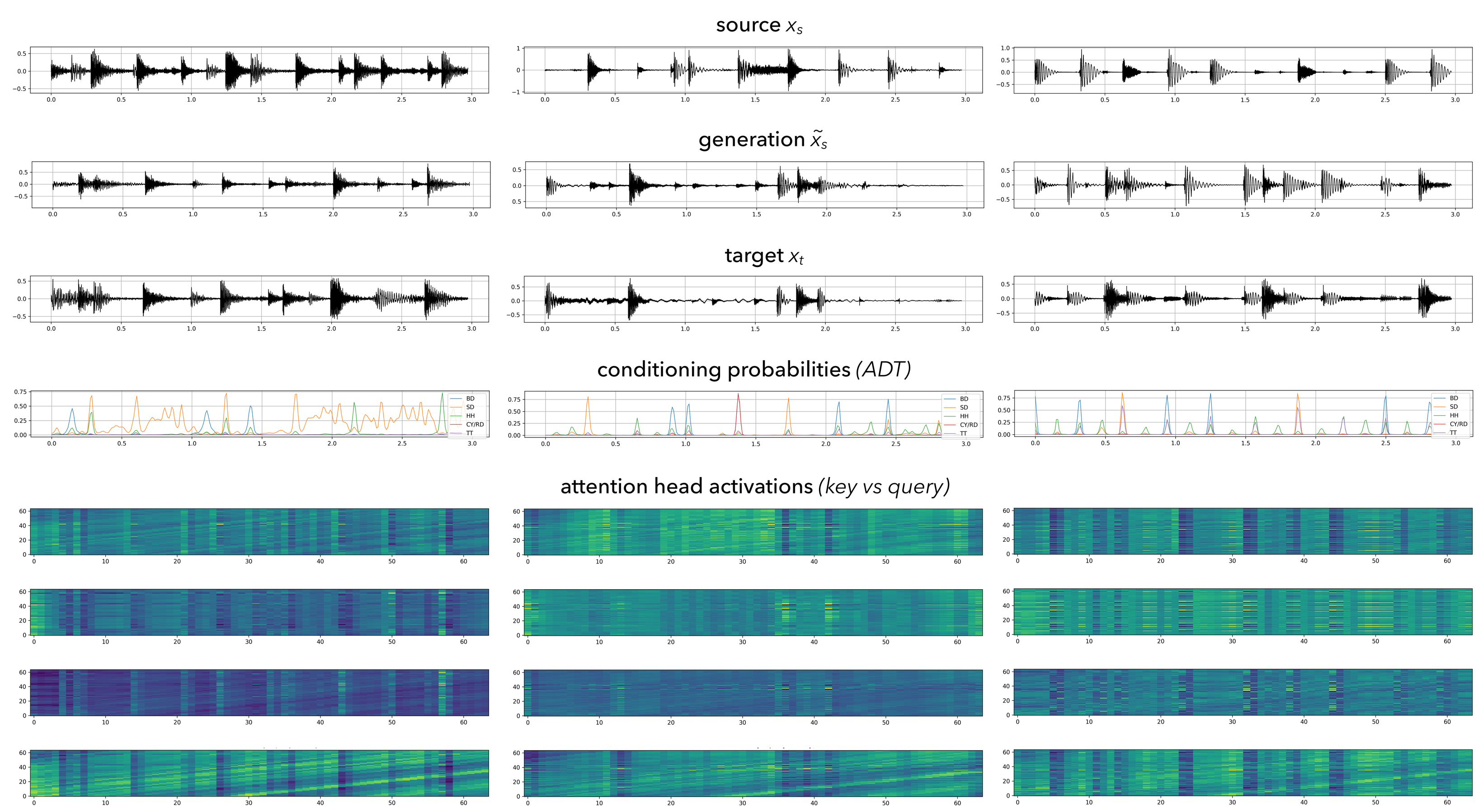

Figure 4: Examples of rhythm transformation, with source x_s, generation $\tilde{x}_s$, and target x_t waveforms, target conditioning probabilities (ADT), and attention head activations.

Audio examples: https://jhockman.github.io/

SODA SCHOOL OF DIGITAL ARTS

Manchester Metropolitan University

beatoven.ai